Simple Bounds for Noisy Linear Inverse Problems with Exact Side Information

Samet Oymak Christos Thrampoulidis Babak Hassibi

Department of Electrical Engineering
Caltech, Pasadena – 91125

Abstract

This paper considers the linear inverse problem where we wish to estimate a structured signal $\mathbf{x}_0$ from its corrupted observations. When the problem is ill-posed, it is natural to associate a convex function $f(\cdot)$ with the structure of the signal. For example, ℓ_1 norm can be used for sparse signals. To carry out the estimation, we consider two well-known convex programs: 1) Second order cone program (SOCP), and, 2) Lasso. Assuming Gaussian measurements, we show that, if precise information about the value $f(\mathbf{x}_0)$ or the ℓ_2-norm of the noise is available, one can do a particularly good job at estimation. In particular, the reconstruction error becomes proportional to the "sparsity" of the signal rather than to the ambient dimension of the noise vector. We connect our results to the existing literature and provide a discussion on their relation to the standard least-squares problem. Our error bounds are non-asymptotic and sharp, they apply to arbitrary convex functions and do not assume any distribution on the noise.

Keywords: *sparse estimation, convex optimization, Lasso, structured signals, Gaussian width, model selection, linear inverse*

1. Introduction

Second order cone programming (SOCP) and the Lasso are two common approaches to perform noise robust model fitting. They are often used for sparse approximation when the signal that underlies the observations is known to have few nonzero entries [1–10]. This work considers the abstract model fitting problem where the signal has some sort of structure and we wish to estimate it from corrupted observations. To accomplish this, we use an abstract structure inducing convex function $f(\cdot)$. Let $\mathbf{x}_0 \in \mathbb{R}^n$ be the true signal to be estimated. We observe $\mathbf{y} = \mathbf{A}\mathbf{x}_0 + \mathbf{z}$ where $\mathbf{A} \in \mathbb{R}^{m\times n}$ is the measurement matrix and $\mathbf{z}$ is the noise vector. Let us now introduce the two problems mentioned above, the SOCP and the Lasso.

1.1. Lasso with exact side information

Lasso is introduced by Tibshirani in [1]. The standard Lasso problem solves,

$$\mathbf{x}_L^* = \arg\min_{\mathbf{x}} \lambda f(\mathbf{x}) + \frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|^2. \tag{1.1}$$

In the program above and in the sequel, $\|\cdot\|$ is the ℓ_2-norm. For the sake of this work, we assume that we know *a priori* the value of the structure inducing function $f(\cdot)$ at $\mathbf{x}_0$. Under this information, we can simplify the problem to the following constrained setup,

$$\mathbf{x}_L^* = \arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|^2 \quad \text{subject to} \quad f(\mathbf{x}) \le f(\mathbf{x}_0). \tag{1.2}$$

*Email: {soymak,cthrampo,hassibi}@caltech.edu. This work was supported in part by the National Science Foundation under grants CCF-0729203, CNS-0932428 and CIF-1018927, by the Office of Naval Research under the MURI grant N00014-08-1-0747, and by a grant from Qualcomm Inc.

1.2. SOCP with exact side information

SOCP is the name given to a class of algorithms. For linear inverse problems, a commonly used instance is the following [2],

$$\mathbf{x}_S^* = \arg\min_{\mathbf{x}} f(\mathbf{x}) \quad \text{subject to} \quad \|\mathbf{y} - \mathbf{A}\mathbf{x}\| \leq \delta.$$

Here δ is a known upper bound on the noise level $\|\mathbf{z}\|$. This ensures that the unknown signal $\mathbf{x}_0$ is feasible for the SOCP. In this work, we will assume the exact information of $\|\mathbf{z}\|$ and solve,

$$\mathbf{x}_S^* = \arg\min_{\mathbf{x}} f(\mathbf{x}) \quad \text{subject to} \quad \|\mathbf{y} - \mathbf{A}\mathbf{x}\| \leq \|\mathbf{z}\|. \tag{1.3}$$

In summary,

- Lasso will assume the knowledge about the signal, $f(\mathbf{x}_0)$.
- SOCP will assume the knowledge about the noise, $\|\mathbf{z}\|$.

We additionally assume that the sensing matrix $\mathbf{A}$ has independent zero-mean, $\frac{1}{m}$ variance Gaussian entries. Our main result provides non-asymptotic and sharp upper bounds on the estimation error terms $\|\mathbf{x}_L^* - \mathbf{x}_0\|$ and $\|\mathbf{x}_S^* - \mathbf{x}_0\|$. When $\mathbf{x}_0$ is a sparse vector and if we pick $f(\cdot)$ to be the ℓ_1 norm, it is now well-known that the estimation error can be as small as $\|\mathbf{z}\|$. In this paper, we restrict our attention to problems (1.2) and (1.3) and we try to answer the following three questions:

- Can we generalize the results on ℓ_1 norm to arbitrary convex functions?
- Can we give very sharp bounds with small and accurate constants?
- Can we do these non-asymptotically, i.e., for possibly very small number of measurements and/or sparsity levels?

2. Result

We will first state the general result and will consider specific examples later on. Let us introduce the "Gaussian width" of a set. This concept is crucial for the statement of our results.

Definition 1 (Gaussian width)**.** *Let $\mathcal{C} \in \mathbb{R}^n$ be a nonempty set. The Gaussian width of $\mathcal{C}$ is denoted by $\omega(\mathcal{C})$ and is defined as,*

$$\omega(\mathcal{C}) = \mathbb{E}\left[\sup_{\mathbf{v}\in\mathcal{C}} \langle \mathbf{v}, \mathbf{g}\rangle\right],$$

where $\mathbf{g} \in \mathbb{R}^n$ has independent standard normal entries.

Next, we require the definition of the tangent cone of a function $f(\cdot)$ at some $\mathbf{x} \in \mathbf{R}^n$. For this definition, let cone$(\cdot)$ and Cl$(\cdot)$ return the conic hull and the closure of a set, respectively.

Definition 2 (Tangent cone)**.** *Assume $f(\cdot) : \mathbb{R}^n \to \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^n$. Denote the set of descend directions $\{\mathbf{v} \in \mathbb{R}^n | f(\mathbf{x}+\mathbf{v}) \leq f(\mathbf{x})\}$ by $D_f(\mathbf{x})$. The tangent cone of $f(\cdot)$ at $\mathbf{x}$ is denoted by $T_f(\mathbf{x})$ and defined as,*

$$T_f(\mathbf{x}) := Cl(cone(D_f(\mathbf{x})).$$

Let $\mathcal{B}^{n-1}$ denote the unit ℓ_2-ball in $\mathbb{R}^n$. For convenience, denote

$$\hat{T}_f(\mathbf{x}) := T_f(\mathbf{x}) \cap \mathcal{B}^{n-1}.$$

Finally, given a vector $\mathbf{g} \in \mathbb{R}^d$ with independent standard normal entries, we define $\gamma_d := \mathbb{E}[\|\mathbf{g}\|]$. It is well known that $\gamma_d = \sqrt{2}\frac{\Gamma(\frac{d+1}{2})}{\Gamma(\frac{d}{2})}$ and $\sqrt{d} \geq \gamma_d \geq \frac{d}{\sqrt{d+1}}$ (see [11]). This definition will simplify our notation in what follows. We are now ready to state our main result.

Theorem 1. *Consider the Lasso and SOCP problems in (1.2) and (1.3), respectively. Assume $\mathbf{z} \in \mathbb{R}^m, \mathbf{x}_0 \in \mathbb{R}^n$ are arbitrary and $\mathbf{A} \in \mathbb{R}^{m\times n}$ has independent $\mathcal{N}(0, \frac{1}{m})$ distributed entries. Assume $m \geq 2$ and $0 \leq t \leq \gamma_m - \boldsymbol{\omega}(\hat{T}_f(\mathbf{x}_0))$. Then, with probability, $1 - 6\exp(-\frac{t^2}{26})$, we have,*

$$\bullet \ \|\mathbf{x}_L^* - \mathbf{x}\| \leq \eta(\mathbf{x}_0, t)\|\mathbf{z}\|, \tag{2.1}$$

$$\bullet \ \|\mathbf{x}_S^* - \mathbf{x}\| \leq 2\eta(\mathbf{x}_0, t)\|\mathbf{z}\|, \tag{2.2}$$

where $\eta(\mathbf{x}_0, t) = \frac{\sqrt{m}}{\gamma_{m-1}} \frac{\boldsymbol{\omega}(\hat{T}_f(\mathbf{x}_0))+t}{\gamma_m - \boldsymbol{\omega}(\hat{T}_f(\mathbf{x}_0))-t}$.

Remark 1: Observing that $\gamma_{m-1}\gamma_m = m-1$ and $\gamma_{m-1} \leq \sqrt{m-1}$ leads to the bound, $\eta(\mathbf{x}_0, t) \leq \frac{\sqrt{m}}{\sqrt{m-1}} \frac{\boldsymbol{\omega}(\hat{T}_f(\mathbf{x}_0))+t}{\sqrt{m-1} - \boldsymbol{\omega}(\hat{T}_f(\mathbf{x}_0))-t}$.

Remark 2: In Theorem 1, we require $\gamma_m \geq \boldsymbol{\omega}(\hat{T}_f(\mathbf{x}_0))$. It has been shown that, this is indeed necessary, [17,19]. When $\gamma_m < \boldsymbol{\omega}(\hat{T}_f(\mathbf{x}_0))$, it is futile to expect noise robustness, as one cannot perfectly recover $\mathbf{x}_0$ from *noiseless* observations $\mathbf{y} = \mathbf{A}\mathbf{x}$ (cf. Theorem 3.4 of [15]).

Our bound is *only* in terms of the Gaussian width; which has been the subject of several works [11,15,17,27–29]. This makes it possible to apply Theorem 1 for specific choices of $f(\cdot)$ and $\mathbf{x}_0$ previously studied in the literature.

3. State-of-the-art applications

We will now state our results for specific signal choices by making use of the existing results in the literature that compute upper bounds on the Gaussian width term $\boldsymbol{\omega}(\hat{T}_f(\mathbf{x}_0))$.

• **Sparse signals:** When $\mathbf{x}_0$ is a k-sparse signal and $f(\cdot)$ is the ℓ_1 norm, we have $\boldsymbol{\omega}(\hat{T}_f(\mathbf{x}_0)) \leq \sqrt{2k\log\frac{2n}{k}}$, [11]. Hence, we have the following.

Corollary 1. *Suppose $\mathbf{x}_0$ is a k-sparse signal and $0 \leq t \leq \sqrt{m-1} - \sqrt{2k\log\frac{2n}{k}}$. Pick $f(\cdot)$ to be the ℓ_1 norm. Then, with probability $1 - 6\exp(-\frac{t^2}{26})$,*

$$\|\mathbf{x}_L^* - \mathbf{x}_0\| \leq \|\mathbf{z}\| \frac{\sqrt{m}}{\sqrt{m-1}} \frac{\sqrt{2k\log\frac{2n}{k}} + t}{\sqrt{m-1} - \sqrt{2k\log\frac{2n}{k}} - t}.$$

• **Low-rank matrices:** Nuclear norm (sum of the singular values) is the standard choice to encourage a low-rank solution. Suppose $\mathbf{x}_0$ is a rank-r matrix of size $d \times d$. For this choice, it is known that $\boldsymbol{\omega}(\hat{T}_f(\mathbf{x}_0)) \leq \sqrt{3r(2d-r)}$, [11].

Corollary 2. *Suppose $\mathbf{x}_0 \in \mathbb{R}^{d\times d}$ is a rank-r matrix and $0 \leq t \leq \sqrt{m-1} - \sqrt{3r(2d-r)}$. Pick $f(\cdot)$ to be the nuclear norm. Then, with probability $1 - 6\exp(-\frac{t^2}{26})$,*

$$\|\mathbf{x}_L^* - \mathbf{x}_0\| \leq \|\mathbf{z}\| \frac{\sqrt{m}}{\sqrt{m-1}} \frac{\sqrt{3r(2d-r)} + t}{\sqrt{m-1} - \sqrt{3r(2d-r)} - t}.$$

• **Block-sparse signals:** Suppose the entries of $\mathbf{x}_0$ can be partitioned into q known blocks of size b and only k of these q blocks are nonzero. The standard function to encourage block-sparsity is the $\ell_{1,2}$ norm, which sums the ℓ_2 norms of the individual blocks. For this choice, it is known that $\boldsymbol{\omega}(\hat{T}_f(\mathbf{x}_0)) \leq \sqrt{4k(b + \log\frac{q}{k})}$, [27].

Corollary 3. *Suppose $\mathbf{x}_0 \in \mathbb{R}^{qb}$ is a k-block-sparse signal and $0 \leq t \leq \sqrt{m-1} - \sqrt{4k(b+\log\frac{q}{k})}$. Pick $f(\cdot)$ to be the $\ell_{1,2}$ norm. Then, with probability $1 - 6\exp(-\frac{t^2}{26})$,*

$$\|\mathbf{x}_L^* - \mathbf{x}_0\| \leq \|\mathbf{z}\| \frac{\sqrt{m}}{\sqrt{m-1}} \frac{\sqrt{4k(b+\log\frac{q}{k})} + t}{\sqrt{m-1} - \sqrt{4k(b+\log\frac{q}{k})} - t}.$$

• **Other low-dimensional models**: There are increasingly more signal classes that exhibit low-dimensionality and to which our results would apply. Some of these are as follows.

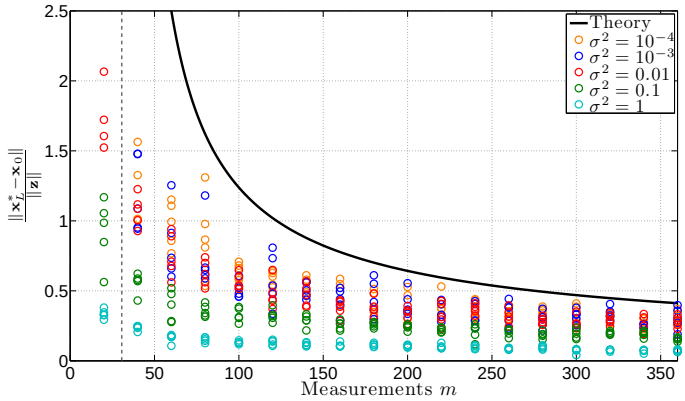

Figure 1: We considered the sparse signal recovery setup of Corollary 1. We set $n = 500$, $k = 5$ and varied m from 0 to 360. Nonzero entries of $\mathbf{x}_0$ is generated with $\mathcal{N}(0,1)$ and then normalized to ensure unit norm. $\mathbf{z}$ and $\mathbf{A}$ has $\mathcal{N}(0,\sigma^2)$ and $\mathcal{N}(0,\frac{1}{m})$ entries respectively. Dashed line corresponds to the phase transition line $m = \boldsymbol{\omega}(\hat{T}_f(\mathbf{x}_0))^2$.

- Non-negativity constraint: $\mathbf{x}_0$ has non-negative entries, [34].
- Low-rank plus sparse matrices: $\mathbf{x}_0$ can be represented as sum of a low-rank and a sparse matrix, [33].
- Signals with sparse gradient: Rather than $\mathbf{x}_0$, its gradient $\mathbf{d}_{\mathbf{x}_0}(i) = \mathbf{x}_0(i) - \mathbf{x}_0(i-1)$ is sparse, [28].
- Low-rank tensors: $\mathbf{x}_0$ is a tensor and its unfoldings are low-rank matrices (see [29,30]).
- Simultaneously sparse and low-rank matrices: For instance, $\mathbf{x}_0 = \mathbf{s}\mathbf{s}^T$ for a sparse vector $\mathbf{s}$, [31,32].

For more examples, the reader is referred to [11,13,15,17].

4. Interpretation of the results

We will now argue that, one can easily interpret our results when the system $\mathbf{y} = \mathbf{A}\mathbf{x}_0 + \mathbf{z}$ is seen as an $m \times \boldsymbol{\omega}(\hat{T}_f(\mathbf{x}_0))^2$ system rather than $m \times n$.

4.1. Comparison to least squares

Consider the least-squares problem where one simply solves,

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|. \tag{4.1}$$

It is clear that when $m < n$, (4.1) is hopeless and when $m > n$ and $\mathbf{A}$ has i.i.d. entries, $\mathbf{A}$ becomes full rank and the solution is $\mathbf{x}^* = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{y}$. Hence, denoting the projection of $\mathbf{z}$ onto the range space of $\mathbf{A}$ by $\text{Proj}(\mathbf{z}, \text{Range}(\mathbf{A}))$ and the minimum singular value of $\mathbf{A}$ by $\sigma_{min}(\mathbf{A})$,

$$\|\mathbf{x}^* - \mathbf{x}_0\|^2 = \mathbf{z}^T\mathbf{A}(\mathbf{A}^T\mathbf{A})^{-2}\mathbf{A}^T\mathbf{z} \leq \left(\frac{\|\text{Proj}(\mathbf{z}, \text{Range}(\mathbf{A}))\|}{\sigma_{min}(\mathbf{A})}\right)^2, \tag{4.2}$$

It is well known that, when $\mathbf{A}$ has $\mathcal{N}(0,\frac{1}{m})$ entries, $\sigma_{min}(\mathbf{A}) \approx 1 - \sqrt{\frac{n}{m}}$, [38]. Also, since the range space is generated uniformly at random, $\|\text{Proj}(\mathbf{z}, \text{Range}(\mathbf{A}))\| \approx \sqrt{\frac{n}{m}}\|\mathbf{z}\|$. Consequently,

$$\|\mathbf{x}^* - \mathbf{x}_0\| \lesssim \|\mathbf{z}\| \frac{\sqrt{n}}{\sqrt{m} - \sqrt{n}}. \tag{4.3}$$

So, what is the relation between (4.3) and (2.1)? Ignoring the t's and using $\gamma_m \approx \sqrt{m}$ in (2.1) , we find,

$$\|\mathbf{x}_L^* - \mathbf{x}\| \lesssim \|\mathbf{z}\| \frac{\boldsymbol{\omega}(\hat{T}_f(\mathbf{x}_0))}{\sqrt{m} - \boldsymbol{\omega}(\hat{T}_f(\mathbf{x}_0))}. \tag{4.4}$$

One can move from (4.4) to (4.3) by simply replacing the $\boldsymbol{\omega}(\hat{T}_f(\mathbf{x}_0))$ terms with $\sqrt{n}$. This indeed indicates that the Lasso and SOCP problems behave as $m \times \boldsymbol{\omega}(\hat{T}_f(\mathbf{x}_0))^2$ systems rather than $m \times n$ ones.

4.2. Comparison to related works

Sparse recovery: A classical result states that, when $\mathbf{x}_0$ is a sparse signal and when $\mathbf{A}$ has independent $\mathcal{N}(0, \frac{1}{m})$ entries the Lasso estimation error obeys $\mathcal{O}\left(\|\mathbf{z}\|\sqrt{\frac{k \log n}{m}}\right)$ when $m = \Omega(k \log \frac{n}{k})$, [3, 13, 18, 21, 26]. Our bound given in Corollary 1 is fully consistent with this, however, we provide very small and accurate constants. In particular, the phase transition occurring around $2k \log \frac{2n}{k}$ number of measurements shows up explicitly in our bound in Corollary 1 (see the term $\sqrt{m-1} - \sqrt{2k \log \frac{2n}{k}}$ in the denominator).

Generalized linear inverse problems: Close to the present paper is the work due to [11]. In [11], Chandrasekaran et al. perform error analysis of the SOCP problem. Their result (cf. Corollary 3.3 in [11]) shows that with probability $1 - \exp(-\frac{1}{2}t^2)$,

$$\|\mathbf{x}_S^* - \mathbf{x}\| \leq 2\sqrt{m} \frac{\|\mathbf{z}\|}{\gamma_m - \boldsymbol{\omega}(\hat{T}_f(\mathbf{x}_0)) - t}. \tag{4.5}$$

Our approach is related; however, we provide a more careful analysis. As a result of this, and in contrast to the error bound in (4.5) which grows linearly with the noise level $\|\mathbf{z}\|$, our bound (2.1) is scaled by a constant factor of $\frac{\boldsymbol{\omega}(\hat{T}_f(\mathbf{x}_0))}{\sqrt{m}}$. This is due to the fact that we are able to carefully remove a significant component of the noise which cannot contribute to the error term.

Sharp error bounds for the Lasso estimator: There has been significant research interest in characterizing the error performance of the Lasso estimators. [13] provides a unified analysis of the error performance of the Lasso estimator (1.1), which can be specialized to many regularizer functions. More recent works establish *sharper* bounds for the Lasso estimation error. In [9,10,19]; Bayati, Montanari and Donoho provide explicit characterizations in an *asymptotic* setting for $f(\cdot) = \|\cdot\|_1$. Closer in nature to the present paper, are the works [16] and [15]. The author in [16] analyzes the Lasso problem (1.2) with prior information on $f(\mathbf{x}_0)$ when $f(\cdot) = \|\cdot\|_1$. [15] generalizes the *precise* analysis to arbitrary convex functions and, most importantly, extends it to penalized Lasso problems of the form (1.1). Although tighter, the bounds in [15] require stronger assumptions than ours, namely, an i.i.d. Gaussian noise vector $\mathbf{z}$ and an asymptotic setting where m and $\boldsymbol{\omega}(\hat{T}_f(\mathbf{x}_0))$ is large enough. Their results translates to our framework as,

$$\|\mathbf{x}_L^* - \mathbf{x}\| \lesssim \|\mathbf{z}\| \frac{\boldsymbol{\omega}(\hat{T}_f(\mathbf{x}_0))}{\sqrt{m - \boldsymbol{\omega}(\hat{T}_f(\mathbf{x}_0))^2}}. \tag{4.6}$$

The difference between (4.4) and (4.6) is in the denominator. $\sqrt{m - \boldsymbol{\omega}(\hat{T}_f(\mathbf{x}_0))^2} \geq \sqrt{m} - \boldsymbol{\omega}(\hat{T}_f(\mathbf{x}_0))$ for all regimes of $0 \leq \boldsymbol{\omega}(\hat{T}_f(\mathbf{x}_0))^2 < m$. The contrast becomes significant when $m \approx \boldsymbol{\omega}(\hat{T}_f(\mathbf{x}_0))^2$. In particular, setting $m = (1+\epsilon)^2 \boldsymbol{\omega}(\hat{T}_f(\mathbf{x}_0))^2$, we have,

$$\frac{\sqrt{m - \boldsymbol{\omega}(\hat{T}_f(\mathbf{x}_0))^2}}{\sqrt{m} - \boldsymbol{\omega}(\hat{T}_f(\mathbf{x}_0))} = \frac{\sqrt{2\epsilon + \epsilon^2}}{\epsilon} = \sqrt{\frac{2}{\epsilon} + 1}.$$

In summary, when ϵ is large, the bounds of this paper are as good as those of [9, 10, 15, 16]. When ϵ is small, they can be arbitrarily worse. Simulation results (see Figure 1) verify that the error bounds of Theorem 1 become sharp for large number of measurements m. This difference can be intuitively explained by considering the least-squares

error in (4.2). There, using $\sigma_{min}(\mathbf{A})$ as an upper bound results in a looser bound. For a vector $\mathbf{z}$ independent of $\mathbf{A}$, we actually have,

$$\frac{n}{m-n}\|\mathbf{z}\|^2 \approx \mathbf{z}^T\mathbf{A}(\mathbf{A}^T\mathbf{A})^{-2}\mathbf{A}^T\mathbf{z} < \left(\frac{\|\text{Proj}(\mathbf{z}, \text{Range}(\mathbf{A}))\|}{\sigma_{min}(\mathbf{A})}\right)^2 \approx \left(\frac{\sqrt{n}}{\sqrt{m}-\sqrt{n}}\|\mathbf{z}\|\right)^2.$$

In this sense, [15] considers the precise behavior of the left-hand side in (4.2) and we consider the looser bound given in the right-hand side; which makes use of the minimum singular value $\sigma_{min}(\mathbf{A})$.

5. Further remarks

5.1. Do we need the exact side information?

For our results, we either assumed knowledge about the signal $f(\mathbf{x}_0)$, or knowledge about the noise $\|\mathbf{z}\|$. It is desirable to not be dependent on such quantities. A natural way to break this dependence is by using the following program,

$$\min_{\mathbf{x}} \lambda f(\mathbf{x}) + \frac{1}{2}\|\mathbf{y}-\mathbf{A}\mathbf{x}\|^2. \tag{5.1}$$

When $f(\cdot)$ is the ℓ_1 norm and $\mathbf{x}_0$ is a sparse signal, the problem becomes the original Lasso program introduced by [1] and it has been analyzed in great depth [1,3–6].Closer to us, Bayati and Montanari analyzes the precise noise characteristics of (5.1) in [9,10]. Analysis of (5.1) for the block-sparse signals and low rank matrices can be found in [23,25]. However, to the best of our knowledge, the existing guarantees are optimal up to a constant; while our bounds are almost exact.

While we leave the analysis of (5.1) to a future work, we should emphasize that, [15] proposed using,

$$\lambda = \frac{\|\mathbf{z}\|}{\sqrt{m}}\tau^*\sqrt{1-\frac{\omega(\hat{T}_f(\mathbf{x}_0))^2}{m}},$$

as the penalty parameter in (5.1) and argued (non rigorously) that (5.1) performs as good as (1.2) with this choice. Here $\tau^* = \arg\min_{\tau\geq 0}\mathbb{E}[\text{dist}(\mathbf{g}, \tau\partial f(\mathbf{x}_0))^2]$ where $\mathbf{g}\sim\mathcal{N}(0,\mathbf{I}_n)$ and $\text{dist}(\mathbf{g}, \tau\partial f(\mathbf{x}_0))$ is the ℓ_2-distance of the vector $\mathbf{g}$ to the τ-scaled subdifferential $\tau\partial f(\mathbf{x}_0)$. Similar choices has been proposed by various works for sparse recovery, [3,5,9,10]. For sparse signals or low-rank matrices, τ^* only depends on sparsity (or rank) of the signal and has been the topic of several works [9–11,15,17,27].

5.2. Adversarial noise

We will now consider the scenario where one has adversarial noise, i.e., noise has the information of the sensing matrix $\mathbf{A}$ and can adapt itself accordingly. In this case, the reconstruction error can become significantly worse. The following proposition illustrates this for the Lasso problem (1.2).

Proposition 1. *Assume $\mathbf{x}_0$ is not a minimizer of $f(\cdot)$. Then, given $\mathbf{A}\in\mathbb{R}^{m\times n}$ with independent $\mathcal{N}(0,\frac{1}{m})$ entries, with probability $1-\exp(-\frac{t^2}{2})$, there exists a noise vector $\mathbf{z}\in\mathbb{R}^m$ and Lasso optimum $\mathbf{x}_L^*$ such that,*

$$\|\mathbf{x}_L^*-\mathbf{x}_0\| \geq \frac{\sqrt{m}}{\gamma_m+t}\|\mathbf{z}\|$$

Proof. Let $\mathbf{x}^* = \arg\min f(\mathbf{x})$. Then, choose $\mathbf{z} = \mathbf{A}(\mathbf{x}^*-\mathbf{x}_0)$; which yields $\mathbf{y} = \mathbf{A}\mathbf{x}_0+\mathbf{z} = \mathbf{A}\mathbf{x}^*$. By construction, $\mathbf{z}\sim\mathcal{N}(0,\frac{\|\mathbf{x}^*-\mathbf{x}_0\|^2}{m}\mathbf{I}_m)$, hence with probability $1-\exp(-\frac{t^2}{2})$, $\|\mathbf{z}\|\leq(\gamma_m+t)\frac{\|\mathbf{x}^*-\mathbf{x}_0\|}{\sqrt{m}}$. Since $f(\mathbf{x}^*)\leq f(\mathbf{x}_0)$ and $\mathbf{A}\mathbf{x}^*-\mathbf{y}=0$, $\mathbf{x}^*$ is a (feasible) minimizer of (1.2) and $\|\mathbf{x}^*-\mathbf{x}_0\|\geq\frac{\sqrt{m}}{\gamma_m+t}\|\mathbf{z}\|$. □

Proposition 1 suggests that we can make error as big as the noise term $\|\mathbf{z}\|$. This contrasts with Theorem 1 where the error is approximately $\frac{\omega(\hat{T}_f(\mathbf{x}_0))}{\sqrt{m}}\|\mathbf{z}\|$ for sufficiently large m. The adversarial noise scenario can again be connected to least-squares in Section 4.1. In (4.2), if the noise $\mathbf{z}$ already lies on $\text{Range}(\mathbf{A})$, we will not have the reduction of

$\sqrt{\frac{n}{m}}$ in the error. Similarly, Proposition 1 constructs a noise vector that that lies in Range($\mathbf{A}$) and originates from the tangent cone element $\mathbf{x}^* - \mathbf{x}_0$. Hence, the resulting error norm is amplified by approximately $\frac{\sqrt{m}}{\boldsymbol{\omega}(\hat{T}_f(\mathbf{x}_0))}$.

Our next result gives an upper bound on the worst case error, which is close to the lower bound when $\boldsymbol{\omega}(\hat{T}_f(\mathbf{x}_0)) \ll \gamma_m$. This uses a very similar argument to Corollary 3.3 of [11].

Proposition 2. *Assume $\mathbf{A} \in \mathbb{R}^{m\times n}$ has independent $\mathcal{N}(0, \frac{1}{m})$ entries and assume $t < \gamma_m - \boldsymbol{\omega}(\hat{T}_f(\mathbf{x}_0))$. Then, with probability $1 - \exp(-\frac{t^2}{2})$, the following bound uniformly hold for all noise vectors $\mathbf{z} \in \mathbb{R}^m$.*

$$\max\{\|\mathbf{x}_L^*(\mathbf{z}) - \mathbf{x}_0\|, \|\mathbf{x}_S^*(\mathbf{z}) - \mathbf{x}_0\|\} \leq \frac{2\sqrt{m}\|\mathbf{z}\|}{\gamma_m - \boldsymbol{\omega}(\hat{T}_f(\mathbf{x}_0)) - t}$$

Proof. From Lemma 1, with probability $1 - \exp(-\frac{t^2}{2})$, we have,

$$\min_{\mathbf{v}\in T_f(\mathbf{x}_0)\cap\mathcal{S}^{n-1}} \|\sqrt{m}\mathbf{A}\mathbf{v}\| \geq \gamma_m - \boldsymbol{\omega}(\hat{T}_f(\mathbf{x}_0)) - t$$

Assuming this happens, we will show the result.
Proof for Lasso: $\mathbf{x}_0$ is feasible for (1.2) hence $\|\mathbf{y} - \mathbf{A}\mathbf{x}_L^*\| \leq \|\mathbf{z}\|$. Also $\mathbf{x}_L^* - \mathbf{x}_0 \in T_f(\mathbf{x}_0)$. Consequently,

$$\|\mathbf{x}_L^* - \mathbf{x}_0\|\frac{\gamma_m - \boldsymbol{\omega}(\hat{T}_f(\mathbf{x}_0)) - t}{\sqrt{m}} - \|\mathbf{z}\| \leq \|\mathbf{A}(\mathbf{x}_L^* - \mathbf{x}_0)\| - \|\mathbf{z}\| \leq \|\mathbf{A}(\mathbf{x}_L^* - \mathbf{x}_0) - \mathbf{z}\| = \|\mathbf{A}\mathbf{x}_L^* - \mathbf{y}\| \leq \|\mathbf{A}\mathbf{x}_0 - \mathbf{y}\| = \|\mathbf{z}\|. \tag{5.2}$$

Proof for SOCP: $\mathbf{x}_0$ is feasible for (1.3) hence $f(\mathbf{x}_S^*) \leq f(\mathbf{x}_0)$ and $\|\mathbf{y} - \mathbf{A}\mathbf{x}_S^*\| \leq \|\mathbf{z}\|$ holds. Hence (5.2) will apply for $\mathbf{x}_S^*$ as well.

□

6. Proof of the Main Result

We begin with introducing some necessary notation in Section 6.1. In Section 6.2, we enlist two critical results for our analysis. Finally, Section 6.3 provides the proof.

6.1. Notation

Throughout the proofs, $\mathbf{A}T_f(\mathbf{x}_0)$ will denote the cone obtained by multiplying elements of $T_f(\mathbf{x}_0)$ by $\mathbf{A}$., i.e.,

$$\mathbf{A}T_f(\mathbf{x}_0) = \{\mathbf{A}\mathbf{v} \in \mathbb{R}^m \mid \mathbf{v} \in T_f(\mathbf{x}_0)\}.$$

Let $\mathcal{C} \in \mathbb{R}^n$ be a convex subset of the unit ℓ_2-sphere $\mathcal{S}^{n-1}$. Then, the minimum singular value of $\mathbf{A} \in \mathbb{R}^{m\times n}$ restricted to $\mathcal{C}$ is defined as,

$$\sigma_{min}(\mathbf{A}, \mathcal{C}) = \min_{\mathbf{v}\in\mathcal{C}} \|\mathbf{A}\mathbf{v}\|.$$

Observe that, $\sigma_{min}(\mathbf{A}, \mathcal{S}^{n-1})$ reduces to the standard definition of the minimum singular value of the matrix $\mathbf{A}$. The projection of a vector $\mathbf{v} \in \mathbb{R}^n$ onto a closed and convex set $\mathcal{C}$ is the unique vector $\text{Proj}(\mathbf{v}, \mathcal{C}) = \arg\min_{\mathbf{s}\in\mathcal{C}} \|\mathbf{v} - \mathbf{s}\|$.

When $\mathcal{C}$ is a closed and convex cone, its polar is defined as $\mathcal{C}^\circ = \{\mathbf{u} \big| \mathbf{u}^T\mathbf{v} \leq 0, \text{ for all } \mathbf{v} \in \mathcal{C}\}$. Moreau's Decomposition Theorem [36], says that, any vector $\mathbf{v}$ can be decomposed as,

$$\mathbf{v} = \text{Proj}(\mathbf{v}, \mathcal{C}) + \text{Proj}(\mathbf{v}, \mathcal{C}^\circ), \text{ where } \langle \text{Proj}(\mathbf{v}, \mathcal{C}), \text{Proj}(\mathbf{v}, \mathcal{C}^\circ)\rangle = 0. \tag{6.1}$$

6.2. Preliminary Results

The next lemma is due to Gordon [14] and relates the Gaussian width to the restricted eigenvalue. This concept is similar to restricted isometry property and has been topic of several related papers, [3, 11–13].

Lemma 1 (Restricted eigenvalue). *Let* $\mathbf{G} \in \mathbb{R}^{m\times n}$ *have independent standard normal entries and* $\mathcal{C} \in \mathcal{S}^{n-1}$. *Assume* $0 \le t \le \gamma_m - \boldsymbol{\omega}(\mathcal{C})$. *Then,*

$$\mathbb{P}\left(\min_{\mathbf{v}\in\mathcal{C}} \|\mathbf{G}\mathbf{v}\| \ge \gamma_m - \boldsymbol{\omega}(\mathcal{C}) - t\right) \ge 1 - \exp(-\frac{t^2}{2}).$$

The next theorem is the main technical contribution of this work. It provides an upper bound on the correlation between a vector and elements of a cone multiplied by a Gaussian matrix.

Theorem 2 (Restricted correlation). *Let* $\mathcal{C} \in \mathbb{R}^n$ *be a convex and closed cone,* $m \ge 2$ *and* $\mathbf{z} \in \mathbb{R}^m$ *be arbitrary. Let* $\mathbf{G} \in \mathbb{R}^{m\times n}$ *have independent standard normal entries. For any* $t \ge 0$, *pick* $\alpha \ge \frac{\boldsymbol{\omega}(\hat{T}_f(\mathbf{x}_0))+t}{\gamma_{m-1}}\|\mathbf{z}\|$. *Then,*

$$\sup_{\mathbf{v}\in\mathcal{C}\cap\mathcal{S}^{n-1}} \{\mathbf{z}^T\mathbf{G}\mathbf{v} - \alpha\|\mathbf{G}\mathbf{v}\|\} \le 0, \tag{6.2}$$

with probability $1 - 5\exp(-\frac{t^2}{26})$.

6.3. Proof of Theorem 1

We will start by providing deterministic bounds on the estimation error. Then, with the help of Lemma 1 and Theorem 2, we will finalize the proof.

Lemma 2 (Deterministic error bounds). *Consider the problems* (1.2) *and* (1.3). *We have,*

$$\max\{\|\mathbf{x}_L^* - \mathbf{x}_0\|, \frac{1}{2}\|\mathbf{x}_S^* - \mathbf{x}_0\|\} \le \frac{\|Proj(\mathbf{z}, \mathbf{A}T_f(\mathbf{x}_0))\|}{\sigma_{min}(\mathbf{A}, T_f(\mathbf{x}_0)\cap\mathcal{S}^{n-1})}.$$

Proof. Using (6.1), let us write, $\mathbf{z} = \mathbf{z}_1 + \mathbf{z}_2$ where $\mathbf{z}_1 = \text{Proj}(\mathbf{z}, \mathbf{A}T_f(\mathbf{x}_0))$, $\mathbf{z}_2 = \text{Proj}(\mathbf{z}, (\mathbf{A}T_f(\mathbf{x}_0))^\circ)$, $\mathbf{z}_1^T\mathbf{z}_2 = 0$.

• *Lasso:* Let $\mathbf{w}^* = \mathbf{x}_L^* - \mathbf{x}_0$. We will first show that $\|\mathbf{A}\mathbf{w}^*\| \le \|\mathbf{z}_1\|$. Assume it is *not* the case and let $\mathbf{w}' = \frac{\|\mathbf{z}_1\|}{\|\mathbf{A}\mathbf{w}^*\|}\mathbf{w}^*$. From convexity, $f(\mathbf{x}_0 + \mathbf{w}') \le f(\mathbf{x}_0)$, hence $\mathbf{w}'$ is feasible. We will show that $\|\mathbf{z} - \mathbf{A}\mathbf{w}'\| < \|\mathbf{z} - \mathbf{A}\mathbf{w}^*\|$, which will contradict with the optimality of $\mathbf{w}^*$.

$$\|\mathbf{z} - \mathbf{A}\mathbf{w}^*\|^2 \ge \|\mathbf{z} - \mathbf{A}\mathbf{w}' + \mathbf{A}(\mathbf{w}' - \mathbf{w}^*)\|^2 = \|\mathbf{z} - \mathbf{A}\mathbf{w}'\|^2 + 2\left\langle \mathbf{z} - \mathbf{A}\mathbf{w}', \mathbf{A}(\mathbf{w}' - \mathbf{w}^*)\right\rangle + \|\mathbf{A}(\mathbf{w}' - \mathbf{w}^*)\|^2$$

Now, observe that,

$$\begin{aligned}\left\langle \mathbf{z} - \mathbf{A}\mathbf{w}', \mathbf{A}(\mathbf{w}' - \mathbf{w}^*)\right\rangle &\ge -\|\mathbf{z}_1\|\|\mathbf{A}(\mathbf{w}' - \mathbf{w}^*)\| - \left\langle \mathbf{A}\mathbf{w}', \mathbf{A}(\mathbf{w}' - \mathbf{w}^*)\right\rangle \\ &\ge -\|\mathbf{z}_1\|\|\mathbf{A}(\mathbf{w}' - \mathbf{w}^*)\| + \|\mathbf{A}\mathbf{w}'\|\|\mathbf{A}(\mathbf{w}' - \mathbf{w}^*)\| \\ &\ge (\|\mathbf{A}\mathbf{w}'\| - \|\mathbf{z}_1\|)\|\mathbf{A}(\mathbf{w}' - \mathbf{w})\| > 0\end{aligned}$$

Hence $\|\mathbf{A}\mathbf{w}^*\| \le \|\mathbf{z}_1\|$. To conclude, we use the fact that $\|\mathbf{w}^*\| \le \frac{\|\mathbf{A}\mathbf{w}^*\|}{\sigma_{min}(\mathbf{A}, T_f(\mathbf{x}_0)\cap\mathcal{S}^{n-1})}$.

• *SOCP:* Let $\mathbf{w}^* = \mathbf{x}_S^* - \mathbf{x}_0$. Then, the problem becomes,

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} f(\mathbf{x}_0 + \mathbf{w}) \quad \text{subject to} \quad \|\mathbf{z} - \mathbf{A}\mathbf{w}\| \le \delta$$

First observe that 0 is feasible, hence $\mathbf{w}^* \in T_f(\mathbf{x}_0)$. Then, for any $\mathbf{w} \in T_f(\mathbf{x}_0)$,

$$\|\mathbf{z}\|^2 = \|\mathbf{z} - \mathbf{A}\mathbf{w}\|^2 = \|\mathbf{z}_2 + \mathbf{z}_1 - \mathbf{A}\mathbf{w}\|^2 = \|\mathbf{z}_2\|^2 + \|\mathbf{z}_1 - \mathbf{A}\mathbf{w}\|^2 + 2\left\langle \mathbf{z}_2, \mathbf{z}_1 - \mathbf{A}\mathbf{w}\right\rangle \ge \|\mathbf{z}_2\|^2 + \|\mathbf{z}_1 - \mathbf{A}\mathbf{w}\|^2,$$

where we used the fact that $\mathbf{z}_2^T\mathbf{A}\mathbf{w} \le 0$ as $\mathbf{A}\mathbf{w} \in \mathbf{A}T_f(\mathbf{x}_0)$. Now, using $\mathbf{w}^* \in T_f(\mathbf{x}_0)$, we find,

$$(\|\mathbf{A}\mathbf{w}^*\| - \|\mathbf{z}_1\|)^2 \le \|\mathbf{z}_1\|^2 \implies \|\mathbf{A}\mathbf{w}^*\| \le 2\|\mathbf{z}_1\| \implies \|\mathbf{w}^*\| \le \frac{2\|\mathbf{z}_1\|}{\sigma_{min}(\mathbf{A}, T_f(\mathbf{x}_0)\cap\mathcal{S}^{n-1})}$$

□

We are now ready to prove Theorem 1

Proof of Theorem 1. Suppose $0 \leq t < \gamma_m - \boldsymbol{\omega}(\hat{T}_f(\mathbf{x}_0))$. We will make use of the fact that, the following events hold with probability $1 - \exp(-\frac{t^2}{2}) - 5\exp(-\frac{t^2}{26})$.

- Observe that $\boldsymbol{\omega}(T_f(\mathbf{x}_0) \cap \mathcal{S}^{n-1}) \leq \boldsymbol{\omega}(\hat{T}_f(\mathbf{x}_0))$. Hence, applying Lemma 1 with $\mathbf{G} = \sqrt{m}\mathbf{A}$ and $\mathcal{C} = T_f(\mathbf{x}_0) \cap \mathcal{S}^{n-1}$, with probability $1 - \exp(-\frac{t^2}{2})$, we have,

$$\sigma_{min}(\mathbf{A}, T_f(\mathbf{x}_0) \cap \mathcal{S}^{n-1}) \geq \frac{\gamma_m - \boldsymbol{\omega}(\hat{T}_f(\mathbf{x}_0)) - t}{\sqrt{m}}. \tag{6.3}$$

- Applying Theorem 2 with $\mathbf{A} = \frac{\mathbf{G}}{\sqrt{m}}$ and $\mathcal{C} = T_f(\mathbf{x}_0)$, with probability $1 - 5\exp(-\frac{t^2}{26})$,

$$\|\text{Proj}(\mathbf{z}, \mathbf{A}T_f(\mathbf{x}_0))\| \leq \frac{\boldsymbol{\omega}(\hat{T}_f(\mathbf{x}_0)) + t}{\gamma_{m-1}}\|\mathbf{z}\|. \tag{6.4}$$

  To see this, pick $\mathbf{v}$ in (6.2) such that $\mathbf{A}\mathbf{v} = \frac{\text{Proj}(\mathbf{z}, \mathbf{A}T_f(\mathbf{x}_0))}{\|\text{Proj}(\mathbf{z}, \mathbf{A}T_f(\mathbf{x}_0))\|}$, which gives $\mathbf{z}^T\mathbf{A}\mathbf{v} = \|\text{Proj}(\mathbf{z}, \mathbf{A}T_f(\mathbf{x}_0))\|$.

Now, the bounds in (2.1) and (2.2) follow when we substitute (6.3) and (6.4) in Lemma 2. □

7. Proof of Theorem 2

7.1. Auxiliary results

There are a few ingredients of the proof. First, we require a result, which allows us to compare two Gaussian processes. This result is again due to Gordon (see Lemma 3.1 in [14]). We make use of a slightly modified version of the original lemma, which can be found in [15] (cf. Lemma 5.1).

Lemma 3 (Comparison Lemma, [14])**.** *Let $\mathbf{G} \in \mathbb{R}^{m\times n}$ have independent standard normal entries. Let $\mathbf{h} \sim \mathcal{N}(0, \mathbf{I}_m)$ and $\mathbf{g} \sim \mathcal{N}(0, \mathbf{I}_n)$. Let $\Phi_1 \subset \mathbb{R}^n$ be an arbitrary set and let $\Phi_2 \subset \mathbb{R}^m$ be a compact set. Then,*

$$\mathbb{P}\left(\min_{\mathbf{x}\in\Phi_1}\max_{\mathbf{a}\in\Phi_2} \mathbf{x}^T\mathbf{G}\mathbf{a} \ \geq c\right) \geq 2\mathbb{P}\left(\min_{\mathbf{x}\in\Phi_1}\max_{\mathbf{a}\in\Phi_2} \|\mathbf{x}\|\mathbf{h}^T\mathbf{a} - \|\mathbf{a}\|\mathbf{g}^T\mathbf{x} \ \geq c\right) - 1.$$

A function $f(\cdot) : \mathbb{R}^n \to \mathbb{R}$ is called L-Lipschitz, if for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,

$$|f(\mathbf{x}) - f(\mathbf{y})| \leq L\|\mathbf{x} - \mathbf{y}\|.$$

The next lemma is a standard result on concentration properties of Lipschitz functions of Gaussian vectors, [35].

Lemma 4. *Let $\mathbf{g} \sim \mathcal{N}(0, \mathbf{I}_n)$, $g \sim \mathcal{N}(0, 1)$ and $f(\cdot) : \mathbb{R}^n \to \mathbb{R}$ be an L-Lipschitz function. Then, for $t \geq 0$,*

$$\mathbb{P}(f(\mathbf{g}) - \mathbb{E}[f(\mathbf{g})] \geq t) \leq \exp(-\frac{t^2}{2L^2}),$$
$$\mathbb{P}(f(\mathbf{g}) - \mathbb{E}[f(\mathbf{g})] \leq -t) \leq \exp(-\frac{t^2}{2L^2}).$$
$$\mathbb{P}(g \geq t) \leq \frac{1}{2}\exp(-\frac{t^2}{2})$$

The following lemma provides a useful identity for the projection of a vector onto a cone.

Lemma 5. *Let $\mathcal{C} \subset \mathbb{R}^n$ be a closed and convex cone and $\mathbf{v} \in \mathbb{R}^n$. Then,*

$$\max_{\mathbf{u}\in\mathcal{C}\cap\mathcal{B}^{n-1}} \mathbf{u}^T\mathbf{v} = \|Proj(\mathbf{v}, \mathcal{C})\|.$$

Proof. From (6.1), we have $\mathbf{v} = \text{Proj}(\mathbf{v}, \mathcal{C}) + \text{Proj}(\mathbf{v}, \mathcal{C}^\circ)$, where $\langle \text{Proj}(\mathbf{v}, \mathcal{C}), \text{Proj}(\mathbf{v}, \mathcal{C}^\circ)\rangle = 0$. For any $\mathbf{u} \in \mathcal{C}$, $\mathbf{u}^T\text{Proj}(\mathbf{v}, \mathcal{C}^\circ) \leq 0$, hence, $\mathbf{u}^T\mathbf{v} \leq \mathbf{u}^T\text{Proj}(\mathbf{v}, \mathcal{C})$. Since $\mathbf{u} \in \mathcal{B}^{n-1}$, we further find from the Cauchy-Schwarz inequality that $\mathbf{u}^T\mathbf{v} \leq \|\text{Proj}(\mathbf{v}, \mathcal{C})\|$. On the other hand, picking $\mathbf{u} = \frac{\text{Proj}(\mathbf{v}, \mathcal{C})}{\|\text{Proj}(\mathbf{v}, \mathcal{C})\|} \in \mathcal{C} \cap \mathcal{B}^{n-1}$, achieves $\mathbf{u}^T\mathbf{v} = \|\text{Proj}(\mathbf{v}, \mathcal{C})\|$. □

7.2. Proof

Proof of Theorem 2. When $\mathbf{z} = 0$, the problem is trivial, hence, assume $\mathbf{z} \neq 0$. If $\alpha \geq \|\mathbf{z}\|$, we clearly have,

$$\sup_{\mathbf{v} \in \mathcal{C} \cap \mathcal{S}^{n-1}} \{\mathbf{z}^T \mathbf{G}\mathbf{v} - \alpha \|\mathbf{G}\mathbf{v}\|\} \leq \sup_{\mathbf{v} \in \mathcal{C} \cap \mathcal{S}^{n-1}} \{\|\mathbf{z}\| \|\mathbf{G}\mathbf{v}\| - \alpha \|\mathbf{G}\mathbf{v}\|\} \leq 0.$$

Hence, without loss of generality, we may assume $\frac{\boldsymbol{\omega}(\hat{T}_f(\mathbf{x}_0)) + t}{\gamma_{m-1}} \|\mathbf{z}\| \leq \alpha < \|\mathbf{z}\|$ and $t < \gamma_{m-1} - \boldsymbol{\omega}(\hat{T}_f(\mathbf{x}_0))$. Define the set $\mathcal{S}_{\mathbf{z}} = \alpha \mathcal{S}^{m-1} - \mathbf{z}$ and let $\hat{\mathcal{C}} := \mathcal{C} \cap \mathcal{B}^{n-1}$. Under this notation,

$$\min_{\mathbf{v} \in \mathcal{C} \cap \mathcal{S}^{n-1}} \alpha \|\mathbf{G}\mathbf{v}\| - \mathbf{z}^T \mathbf{G}\mathbf{v} = \min_{\mathbf{v} \in \mathcal{C} \cap \mathcal{S}^{n-1}} \max_{\mathbf{u} \in \mathcal{S}_{\mathbf{z}}} \mathbf{u}^T \mathbf{G}\mathbf{v}.$$

With this min max formulation, we can apply Lemma 3 and use the fact that $\|\mathbf{v}\| = 1$ to find,

$$\mathbb{P}\left(\min_{\mathbf{v} \in \mathcal{C} \cap \mathcal{S}^{n-1}} \max_{\mathbf{u} \in \mathcal{S}_{\mathbf{z}}} \mathbf{u}^T \mathbf{G}\mathbf{v} \geq 0 \right) \geq 2\mathbb{P}\left(\min_{\mathbf{v} \in \mathcal{C} \cap \mathcal{S}^{n-1}} \max_{\mathbf{u} \in \mathcal{S}_{\mathbf{z}}} \mathbf{h}^T \mathbf{u} - \|\mathbf{u}\| \mathbf{g}^T \mathbf{v} \geq 0 \right) - 1, \tag{7.1}$$

where $\mathbf{h} \sim \mathcal{N}(0, \mathbf{I}_m)$ and $\mathbf{g} \sim \mathcal{N}(0, \mathbf{I}_n)$. For the rest of the proof we focus on the analysis of the *simpler* optimization problem on the right hand side of (7.1). Begin by noting that $\mathcal{C} \cap \mathcal{S}^{n-1} \subset \hat{\mathcal{C}}$, hence,

$$\min_{\mathbf{v} \in \mathcal{C} \cap \mathcal{S}^{n-1}} \max_{\mathbf{u} \in \mathcal{S}_{\mathbf{z}}} \mathbf{h}^T \mathbf{u} - \|\mathbf{u}\| \mathbf{g}^T \mathbf{v} \;\geq\; \min_{\mathbf{v} \in \hat{\mathcal{C}}} \max_{\mathbf{u} \in \mathcal{S}_{\mathbf{z}}} \mathbf{h}^T \mathbf{u} - \|\mathbf{u}\| \mathbf{g}^T \mathbf{v}.$$

The only term in which $\mathbf{v}$ appears above is $\mathbf{g}^T \mathbf{v}$. From Lemma 5, $\max_{\mathbf{v} \in \hat{\mathcal{C}}} \mathbf{g}^T \mathbf{v} = \|\mathrm{Proj}(\mathbf{g}, \mathcal{C})\|$. Hence, we find,

$$\min_{\mathbf{v} \in \hat{\mathcal{C}}} \max_{\mathbf{u} \in \mathcal{S}_{\mathbf{z}}} \mathbf{h}^T \mathbf{u} - \|\mathbf{u}\| \mathbf{g}^T \mathbf{v} \;=\; \max_{\mathbf{u} \in \mathcal{S}_{\mathbf{z}}} \left\{ \mathbf{h}^T \mathbf{u} - \|\mathbf{u}\| \|\mathrm{Proj}(\mathbf{g}, \mathcal{C})\| \right\}.$$

Now, we make the change of variable $\mathbf{u} = \alpha \mathbf{a} - \mathbf{z}$ and write the right-hand side above as,

$$\max_{\mathbf{u} \in \mathcal{S}_{\mathbf{z}}} \left\{ \mathbf{h}^T \mathbf{u} - \|\mathbf{u}\| \|\mathrm{Proj}(\mathbf{g}, \mathcal{C})\| \right\} = \max_{\mathbf{a} \in \mathcal{S}^{m-1}} \left\{ \mathbf{h}^T (\alpha \mathbf{a} - \mathbf{z}) - \|\alpha \mathbf{a} - \mathbf{z}\| \|\mathrm{Proj}(\mathbf{g}, \mathcal{C})\| \right\} \tag{7.2}$$

Recall from (7.1), that we want to lower bound the optimization problem above. The choice of $\mathbf{a}$ is up to us and a good choice will guarantee a good lower bound on the right hand side of (7.2). Let $\hat{\mathbf{z}} := \frac{\mathbf{z}}{\|\mathbf{z}\|}$. Further, denote the projection of $\mathbf{h}$ onto $\mathbf{z}$ as $\mathbf{h}_2 := \hat{\mathbf{z}} \hat{\mathbf{z}}^T \mathbf{h}$. Also, $\mathbf{h}_1 := \mathbf{h} - \mathbf{h}_2$ and $\mathbf{h}_1$ is, by construction, orthogonal to $\mathbf{z}$ and is independent of $\mathbf{h}_2$. Let us choose

$$\mathbf{a} = \sqrt{1 - \beta^2} \frac{\mathbf{h}_1}{\|\mathbf{h}_1\|} + \beta \hat{\mathbf{z}},$$

with

$$\beta = \min \left\{ \frac{\|\mathrm{Proj}(\mathbf{g}, \mathcal{C})\|}{\|\mathbf{h}_1\|}, 1 \right\}.$$

Then,

$$\|\alpha \mathbf{a} - \mathbf{z}\|^2 = \|\alpha \sqrt{1 - \beta^2} \frac{\mathbf{h}_1}{\|\mathbf{h}_1\|} - \hat{\mathbf{z}}(\|\mathbf{z}\| - \alpha \beta)\|^2 = \alpha^2 (1 - \beta^2) + (\|\mathbf{z}\| - \alpha \beta)^2,$$

and, denote,

$$\kappa_1 := \sqrt{\alpha^2 (1 - \beta^2) + (\|\mathbf{z}\| - \alpha \beta)^2} \|\mathrm{Proj}(\mathbf{g}, \mathcal{C})\|$$

Similarly,

$$(\alpha \mathbf{a} - \mathbf{z})^T \mathbf{h} = \alpha \sqrt{1 - \beta^2} \|\mathbf{h}_1\| - (\|\mathbf{z}\| - \alpha \beta) \mathbf{h}_2^T \hat{\mathbf{z}},$$

and, denote,

$$\kappa_2 := \alpha \sqrt{1 - \beta^2} \|\mathbf{h}_1\| \quad \text{and} \quad \kappa_3 := (\|\mathbf{z}\| - \alpha \beta) \mathbf{h}_2^T \hat{\mathbf{z}}.$$

Note that $\mathbf{g}, \mathbf{h}_1, \mathbf{h}_2$ are all independent of each other and individually appears in $\kappa_1, \kappa_2, \kappa_3$ respectively. From (7.2), we are interested in $\mathbb{P}(\kappa_2 - \kappa_1 - \kappa_3 \geq 0)$. Let $\hat{\alpha} = \frac{\alpha}{\|\mathbf{z}\|}$. To lower bound this, we will consider the events,

$$\|\mathbf{h}_1\| \geq \gamma_{m-1} - \tau; \quad \|\text{Proj}(\mathbf{g}, \mathcal{C})\| \leq \boldsymbol{\omega}(\mathcal{C} \cap \mathcal{B}^{n-1}) + \tau; \quad \mathbf{h}_2^T \hat{\mathbf{z}} \leq \tau \tag{7.3}$$

for some $\tau > 0$ (to be determined) and the associated probabilities which are obtained as an application of Lemma 4.

- $\mathbb{P}(\|\mathbf{h}_1\| \geq \gamma_{m-1} - \tau) \geq 1 - \exp(-\frac{\tau^2}{2})$.
- $\mathbb{P}(\|\text{Proj}(\mathbf{g}, \mathcal{C})\| \leq \boldsymbol{\omega}(\mathcal{C} \cap \mathcal{B}^{n-1}) + \tau) \geq 1 - \exp(-\frac{\tau^2}{2})$.
- $\mathbb{P}(\mathbf{h}_2^T \hat{\mathbf{z}} \leq \tau) \geq 1 - \frac{1}{2}\exp(-\frac{\tau^2}{2})$

The first one holds from the fact that ℓ_2-norm is 1-Lipschitz. Second one follows from 1-Lipschitzness of distance to a convex set [37]. Finally, the third bound follows from the fact that $\mathbf{h}_2^T \hat{\mathbf{z}}$ is statistically identical to $\mathcal{N}(0,1)$.

From the initial assumptions $\gamma_{m-1} - \boldsymbol{\omega}(\mathcal{C} \cap \mathcal{B}^{n-1}) > t$. For the rest of the discussion, let $\tau = \frac{t}{3.6}$ and assume the three events in (7.3) hold, which happens with probability $1 - \frac{5}{2}\exp(-\frac{\tau^2}{2}) \geq 1 - \frac{5}{2}\exp(-\frac{t^2}{26})$. We will now show that $\kappa_2 - \kappa_1 - \kappa_3 \geq 0$. First, observe that, we have the following list of inequalities.

$$\begin{aligned}\beta = \frac{\|\text{Proj}(\mathbf{g}, \mathcal{C})\|}{\|\mathbf{h}_1\|} &\leq \frac{\boldsymbol{\omega}(\mathcal{C} \cap \mathcal{B}^{n-1}) + \tau}{\gamma_{m-1} - \tau} \leq \frac{\boldsymbol{\omega}(\mathcal{C} \cap \mathcal{B}^{n-1}) + 2.6\tau}{\gamma_{m-1} - \tau} \\ &\leq \frac{\boldsymbol{\omega}(\mathcal{C} \cap \mathcal{B}^{n-1}) + 3.6\tau}{\gamma_{m-1}} \leq \frac{\alpha}{\|\mathbf{z}\|} < 1.\end{aligned} \tag{7.4}$$

Also, since $\|\mathbf{h}_1\| \geq \gamma_{m-1} - \tau$,

$$(\frac{\alpha}{\|\mathbf{z}\|} - \beta)\|\mathbf{h}_1\| \geq \boldsymbol{\omega}(\mathcal{C} \cap \mathcal{B}^{n-1}) + 2.6\tau - \|\text{Proj}(\mathbf{g}, \mathcal{C})\| \geq 1.6\tau \tag{7.5}$$

Let us focus on $\kappa_2 - \kappa_1$ and let $\hat{\alpha} = \frac{\alpha}{\|\mathbf{z}\|}$. We may write,

$$\frac{\kappa_2 - \kappa_1}{\|\mathbf{z}\|} = \hat{\alpha}\sqrt{1-\beta^2}\|\mathbf{h}_1\| - \sqrt{\hat{\alpha}^2(1-\beta^2) + (1-\hat{\alpha}\beta)^2}\|\text{Proj}(\mathbf{z}, \mathcal{C})\|$$

Further normalizing by $\|\mathbf{h}_1\|$, we find,

$$\hat{\kappa}(\hat{\alpha}) := \frac{\kappa_2 - \kappa_1}{\|\mathbf{z}\|\|\mathbf{h}_1\|} = \hat{\alpha}\sqrt{1-\beta^2} - \beta\sqrt{\hat{\alpha}^2(1-\beta^2) + (1-\hat{\alpha}\beta)^2}$$

Expanding $\hat{\kappa}(\hat{\alpha})$,

$$\hat{\kappa}(\hat{\alpha}) = \hat{\alpha}\sqrt{1-\beta^2} - \beta\sqrt{1 + \hat{\alpha}^2 - 2\hat{\alpha}\beta}$$

For $\hat{\kappa}(\hat{\alpha})$, we have the following result.

Lemma 6. *Let β be same as in (7.4). Then, for $1 \geq \hat{\alpha} \geq \beta$, we have that $\hat{\kappa}(\hat{\alpha}) \geq \sqrt{\frac{1-\beta}{2}}(\hat{\alpha} - \beta)$.*

Proof. Observe that $\hat{\kappa}(\beta) = 0$. Using $0 \leq \beta < 1$ and differentiating with respect to $\hat{\alpha}$, for $\hat{\alpha} \geq \beta$,

$$\hat{\kappa}'(\hat{\alpha}) = \sqrt{1-\beta^2} - \frac{\beta(\hat{\alpha} - \beta)}{\sqrt{1 + \hat{\alpha}^2 - 2\hat{\alpha}\beta}}. \tag{7.6}$$

Differentiating one more time, we find,

$$\hat{\kappa}''(\hat{\alpha}) = -\frac{\beta([1 + \hat{\alpha}^2 - 2\hat{\alpha}\beta] - \hat{\alpha}(\hat{\alpha} - \beta) + \beta(\hat{\alpha} - \beta))}{(1 + \hat{\alpha}^2 - 2\hat{\alpha}\beta)^{3/2}} = \frac{-\beta(1-\beta^2)}{(1 + \hat{\alpha}^2 - 2\hat{\alpha}\beta)^{3/2}} \leq 0.$$

Since the second derivative is nonpositive, this means $\hat{\kappa}'(\hat{\alpha})$ is minimized at $\hat{\alpha} = 1$ over the region $\beta \leq \hat{\alpha} \leq 1$. Consequently, for $1 \geq \hat{\alpha} \geq \beta$, we have,

$$\hat{\kappa}(\hat{\alpha}) \geq \hat{\kappa}(\hat{\alpha}) - \hat{\kappa}(\beta) \geq (\hat{\alpha} - \beta)\hat{\kappa}'(1) \tag{7.7}$$

To find $\hat{\kappa}'(1)$, set $\hat{\alpha} = 1$ in (7.6),

$$\hat{\kappa}'(1) = \sqrt{1-\beta^2} - \frac{\beta(1-\beta)}{\sqrt{2-2\beta}} = \sqrt{1-\beta}(\sqrt{1+\beta} - \frac{\beta}{\sqrt{2}}) \geq \sqrt{\frac{1-\beta}{2}}.$$

Here we used the fact that $\sqrt{1+\beta} - \frac{\beta}{\sqrt{2}}$ is minimized at $\beta = 1$ over $0 \leq \beta \leq 1$, which can be verified by differentiating. Substituting this in (7.7), we find the desired result. □

Now, applying Lemma 6 and using (7.5), we have,

$$\frac{\kappa_2 - \kappa_1}{\|\mathbf{z}\|} = \|\mathbf{h}_1\|(\hat{\kappa}(\hat{\alpha}) - \hat{\kappa}(\beta)) \geq \|\mathbf{h}_1\|(\hat{\alpha} - \beta)\sqrt{\frac{1-\beta}{2}} \geq 1.6\tau\sqrt{\frac{1-\beta}{2}} \tag{7.8}$$

Finally, to bound κ_3, for $1 \geq \hat{\alpha} \geq \beta$, we use $0 \leq \|\mathbf{z}\| - \alpha\beta \leq \|\mathbf{z}\|(1-\beta^2)$. This gives

$$0 \leq \frac{\kappa_3}{\|\mathbf{z}\|} \leq (1-\beta^2)\tau$$

Combining with (7.8), we find,

$$\frac{\kappa_2 - \kappa_1 - \kappa_3}{\|\mathbf{z}\|} \geq 1.6\tau\sqrt{\frac{1-\beta}{2}} - (1-\beta^2)\tau \geq 0$$

Here, the nonnegativity of the right-hand side is equivalent to,

$$2.56\frac{1-\beta}{2} \geq (1-\beta^2)^2 \iff 1.28 \geq (1+\beta)(1-\beta^2)$$

Differentiating the $(1+\beta)(1-\beta^2)$ term, we find that, it is maximized at $\beta = \frac{1}{3}$ and is upper bounded by $\frac{32}{27} \leq 1.28$. In summary, we have shown that, with probability $1 - \frac{5}{2}\exp(-\frac{t^2}{26})$ (7.3) hold with $\tau = \frac{t}{3.6}$, and we have, $\kappa_2 - \kappa_1 - \kappa_3 \geq 0$; which also implies nonnegativity of right-hand side of (7.2). Now, using (7.1), we find the desired result.

□

References

[1] R. Tibshirani, "Regression shrinkage and selection via the lasso." Journal of the Royal Statistical Society, 58:267–288, 1996.

[2] E. J. Candès, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements." Communications on pure and applied mathematics 59.8 (2006): 1207-1223.

[3] P. J. Bickel , Y. Ritov and A. Tsybakov, "Simultaneous analysis of LASSO and Dantzig Selector." The Annals of Statistics, 37(4):1705–1732, 2009.

[4] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp, "Sparsity oracle inequalities for the lasso." Electronic Journal of Statistics, 1:169–194, 2007.

[5] M. J. Wainwright, "Sharp Thresholds for High-Dimensional and Noisy Sparsity Recovery Using ℓ_1-constrained quadratic programming." Information Theory, IEEE Transactions on 55.5 (2009): 2183-2202.

[6] P. Zhao and B. Yu, "On model selection consistency of Lasso." Journal of Machine Learning Research, 7:2541–2567, 2006.

[7] D. L. Donoho, M. Elad, and V. M. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise." IEEE Trans. Info Theory, 52(1):6–18, January 2006.

[8] N. Meinshausen and B. Yu, "Lasso-type recovery of sparse representations for high-dimensional data." Ann. Statist., 37(1):246–270, 2009.

[9] M. Bayati and A. Montanari. "The dynamics of message passing on dense graphs, with applications to compressed sensing." IEEE Transactions on Information Theory, Vol. 57, No. 2, 2011.

[10] M. Bayati and A. Montanari, "The LASSO risk for gaussian matrices." IEEE Transactions on Information Theory, Vol. 58, No. 4, 2012.

[11] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, "The Convex Geometry of Linear Inverse Problems." Foundations of Computational Mathematics. Online First, October 2012.

[12] G. Raskutti, M. J. Wainwright, and B. Yu, "Restricted Eigenvalue Properties for Correlated Gaussian Designs." Journal of Machine Learning Research 11 (2010) 2241-2259.

[13] S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu, "A Unified Framework for High-Dimensional Analysis of M-Estimators with Decomposable Regularizers." Statistical Science 2012, Vol. 27, No. 4, 538–557.

[14] Y. Gordon, "On Milman's inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$." in Geometric Aspects of Functional Analysis, volume 1317 of Lecture Notes in Mathematics, pages 84–106. Springer, 1988.

[15] S. Oymak, C. Thrampoulidis, and B. Hassibi, "The Squared-Error of Generalized LASSO: A Precise Analysis." arXiv:1311.0830.

[16] M. Stojnic, "A framework to characterize performance of LASSO algorithms." arXiv preprint arXiv:1303.7291 (2013).

[17] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp, "Living on the edge: A geometric theory of phase transitions in convex optimization." arXiv:1303.6672.

[18] D. L. Donoho, I. Johnstone, A. Maleki, and A. Montanari, "Compressed sensing overℓ_p balls: Minimax mean square error." Information Theory Proceedings (ISIT), 2011 IEEE International Symposium on. IEEE, 2011.

[19] D. L. Donoho, A. Maleki, and A. Montanari. "The noise-sensitivity phase transition in compressed sensing." Information Theory, IEEE Transactions on 57.10 (2011): 6920-6941.

[20] E. J. Candès and M. A. Davenport, "How well can we estimate a sparse vector?." Applied and Computational Harmonic Analysis 34, 317–323.

[21] E. J. Candès and T. Tao. "The Dantzig selector: Statistical estimation when p is much larger than n." Ann. Stat., 35(6):2313–2351, 2007.

[22] M. S. Lobo, L. Vandenberghe, S. Boyd, and H. Lebret, "Applications of second-order cone programming." Linear algebra and its applications, 284(1), 193–228.

[23] L. Meier, S. van de Geer, and P. Buhlmann, "The group Lasso for logistic regression". J. Roy. Statist. Soc. Ser. B 70 53–71, 2008.

[24] N. Meinshausen and B. Yu. "Lasso-type recovery of sparse representations for high-dimensional data." The Annals of Statistics (2009): 246–270.

[25] V. Koltchinskii, K. Lounici, and A. Tsybakov, "Nuclear norm penalization and optimal rates for noisy matrix completion", Annals of Statistics, 2011.

[26] Belloni, Alexandre, Victor Chernozhukov, and Lie Wang. "Square-root lasso: pivotal recovery of sparse signals via conic programming" Biometrika 98.4 (2011): 791-806.

[27] R. Foygel and L. Mackey, "Corrupted Sensing: Novel Guarantees for Separating Structured Signals." IEEE Transactions on Information Theory. To appear.

[28] J-F. Cai, W. Xu, " Guarantees of Total Variation Minimization for Signal Recovery." arXiv:1301.6791.

[29] C. Mu, B. Huang, J. Wright, and D. Goldfarb, "Square Deal: Lower Bounds and Improved Relaxations for Tensor Recovery." arXiv:1307.5870.

[30] S. Gandy, B. Recht, and I. Yamada."Tensor completion and low-n-rank tensor recovery via convex optimization." Inverse Problems 27.2 (2011): 025010.

[31] S. Oymak, A. Jalali, M. Fazel, Yonina C. Eldar, and B. Hassibi, "Simultaneously Structured Models with Application to Sparse and Low-rank Matrices." arXiv:1212.3753.

[32] E. Richard, P-A. Savalle, and N. Vayatis. "Estimation of simultaneously sparse and low rank matrices." arXiv:1206.6474.

[33] J. Wright, A. Ganesh, K. Min, and Y. Ma, "Compressive principal component pursuit." Information and Inference, 2(1), 32-68.

[34] D. L. Donoho and J. Tanner. "Sparse nonnegative solution of underdetermined linear equations by linear programming." Proceedings of the National Academy of Sciences of the United States of America 102.27 (2005): 9446-9451.

[35] M. Ledoux and M. Talagrand, "Probability in Banach Spaces: Isoperimetry and Processes". Springer, 1991.

[36] J.-J. Moreau, "Fonctions convexes duales et points proximaux dans un espace hilbertien." Note aux C.R. Acad. Sci. Paris 255, 2897–2899 (1962).

[37] D. P. Bertsekas, A. Nedić, and A. E. Ozdaglar. "Convex analysis and optimization." Belmont: Athena Scientific, 2003.

[38] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices." Chapter 5 of: Compressed Sensing, Theory and Applications. Edited by Y. Eldar and G. Kutyniok. Cambridge University Press, 2012.